Reusing raw data for machine learning in MX

Melanie Vollmar

ARISE fellow/EMBL fellow/ Marie Curie fellow

EMBL-EBI

ARISE project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945405.

EMBL

Content

- Problem statement
- Proposed solution
- Contacting users
- User responses
- Likely reason for low response rate
- Raw data usage

Problem statement

Grant proposal for

Analysis-driven data acquisition

“rapid processing and communication to ensure optimal, problem-driven use of synchrotron beamtime”

Idea was to

Develop improved metrics for data utility

Develop feed-forward and feed-backfard information flow to/from structure solution

Develop tools to evaluate and steer crystallographic experiments

Ideally in real time

Based on quality of electron density maps and intermediate success metrics

Suggested solution

Use user data to do data analysis and identify relevant metrics or if necessary create novel metrics

Do this analysis at different key points along the structure solution process, e.g. at data integration, during phasing and when refining

Try out how much data is needed to get reliable results

Check whether these metrics can reliably be used to predict likely structure solution success

Train a machine learning application on reliable metrics to predict the chances of success at different structure solution key points

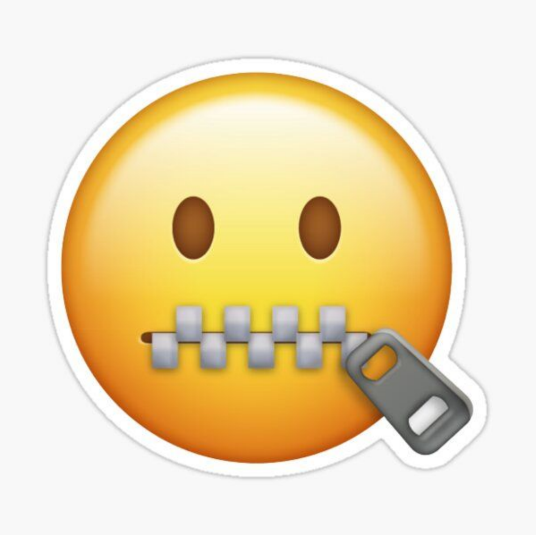
Use the prediction outcome to decide next steps in the analysis workflows

How did we contact users and what did we ask for

- Through CCP4 mailing lists
- Approaching possible users at meetings and conferences
- Asking for help at the end of a presentation

- Structure in the PDB
- Data collected at Diamond
- Which beamline
- What date
- Directory structure and filename
- Image numbers that resulted in a structure
- Willingness to serve as an assessor for the predictions

How did users respond



except for

Arnaud Baslé

Newcastle University

Volunteered;
Served as assessor

~20 datasets

Tobias Krojer

Frank von Delft

SGC

Through personal contacts
Querying their database

~400 datasets

JCSG
Joint Center for Structural Genomics
Developing HT methods for Gene to Structure and Function

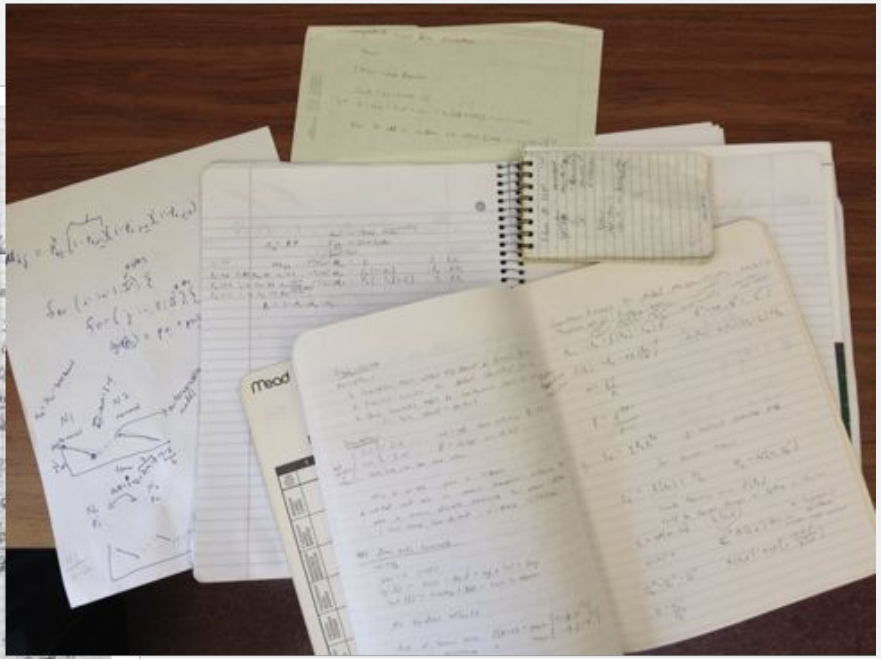
Now at:
https://proteindiffraction.org/

All their data made public
for download

~800 datasets

Likely reason for low response rate

NO data management plan



HDD

Best guess
Data is inaccessible and untraceable

Raw data usage

Training an experimental phasing predictor

JCSG	507 structures
SGC	303 structures
Newcastle	24 structures (independent validation set)

Phasing method:

S/MAD	446 (positive data)
Native	364 (negative data)

Resolution range:
1.05 – 3.8Å

Detector type:
CCD, PAD

X-Ray source:
Synchrotron, in-house

Protein:
6 – 100kDa

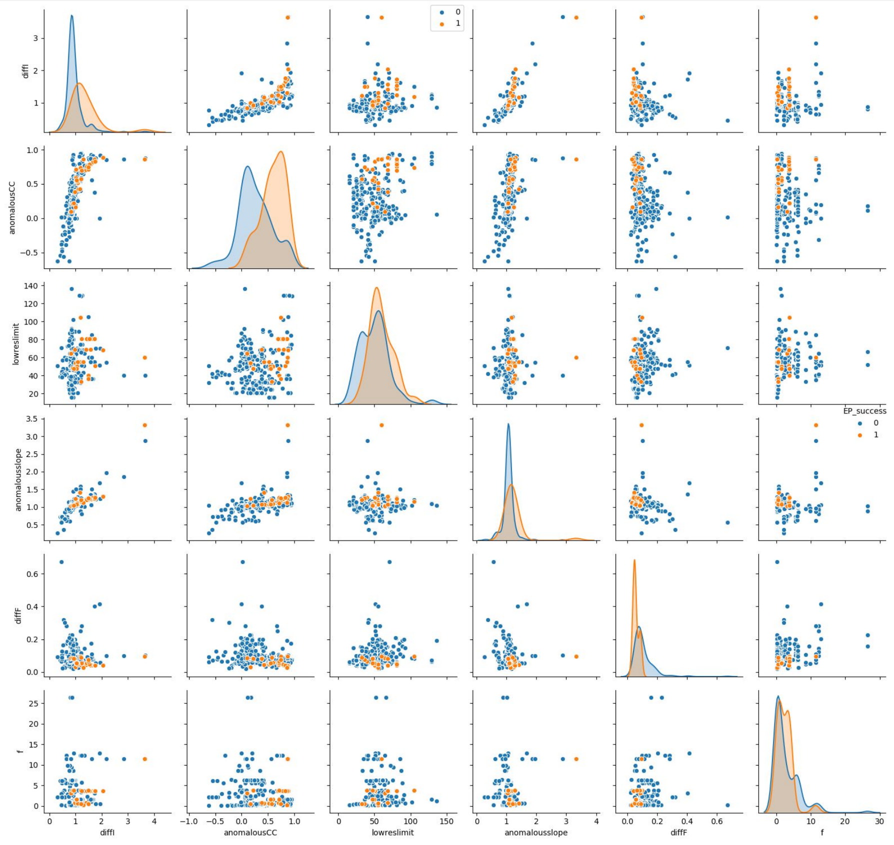
Pre-assessment tried
Linear Pearson’s correlation coefficients
Recursive feature elimination

Classifiers tried
Support vector machine with linear kernel
Support vector machine with RBF kernel
Decision tree
Decision tree with Bagging
Decision tree with AdaBoost
Random forest
Extreme random forest

703 samples (after processing)
stratified test-train split (20/80)
3-fold cross-validation (20/80 split)

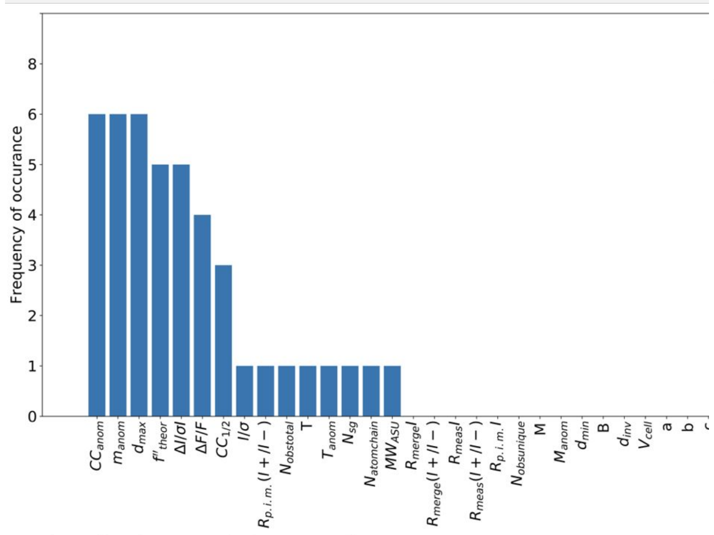
Raw data usage

Training an experimental phasing predictor



3dii/XDS

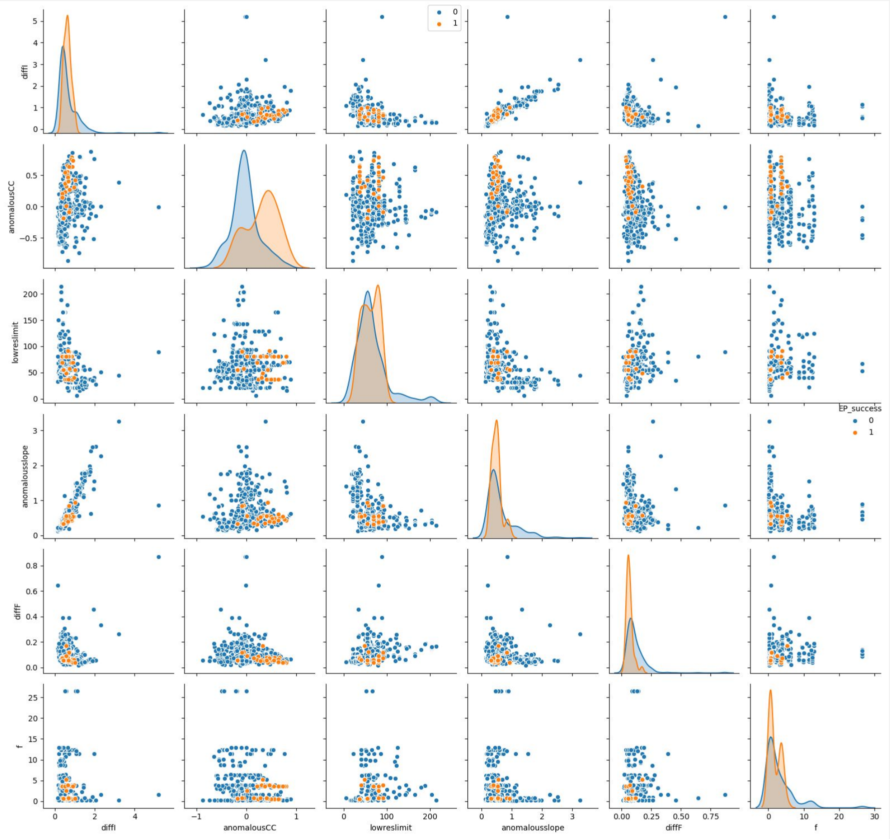
Important decision making features



d_{max} → low resolution cut-off

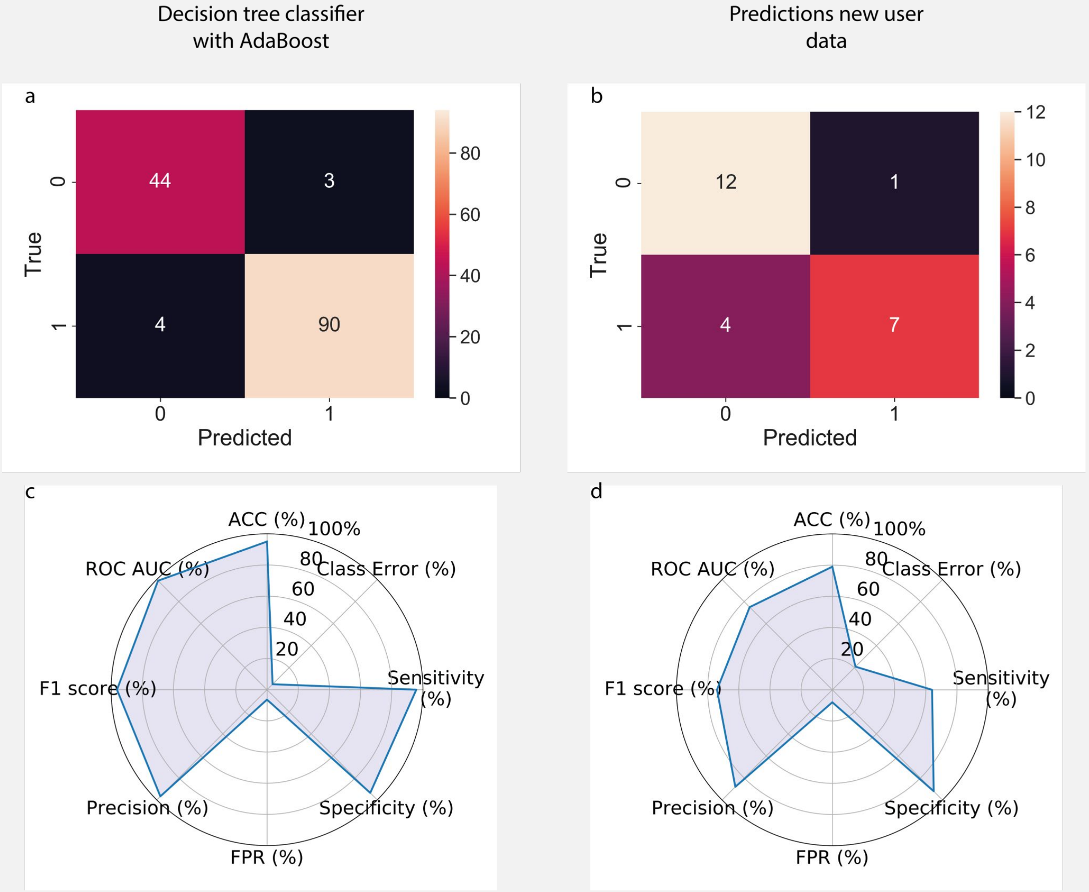
d_{min} → high resolution cut-off

DIALS



Raw data usage

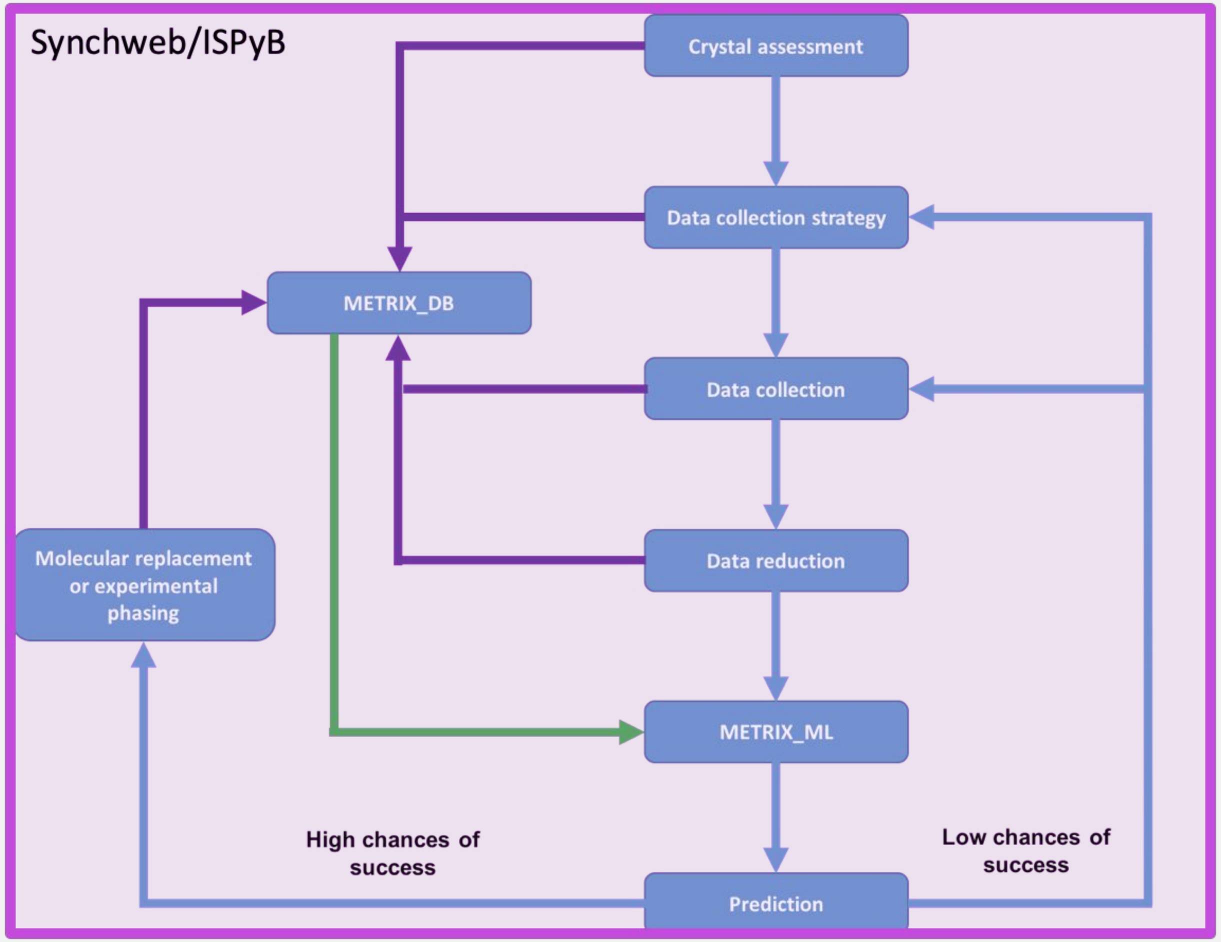
Training an experimental phasing predictor



Decision tree classifier with AdaBoost

Predictions new user data

Vollmar *et al.*, IUCrJ, 2019



Synchweb/ISPyB

Crystal assessment

Data collection strategy

METRIX_DB

Data collection

Data reduction

Molecular replacement or experimental phasing

METRIX_ML

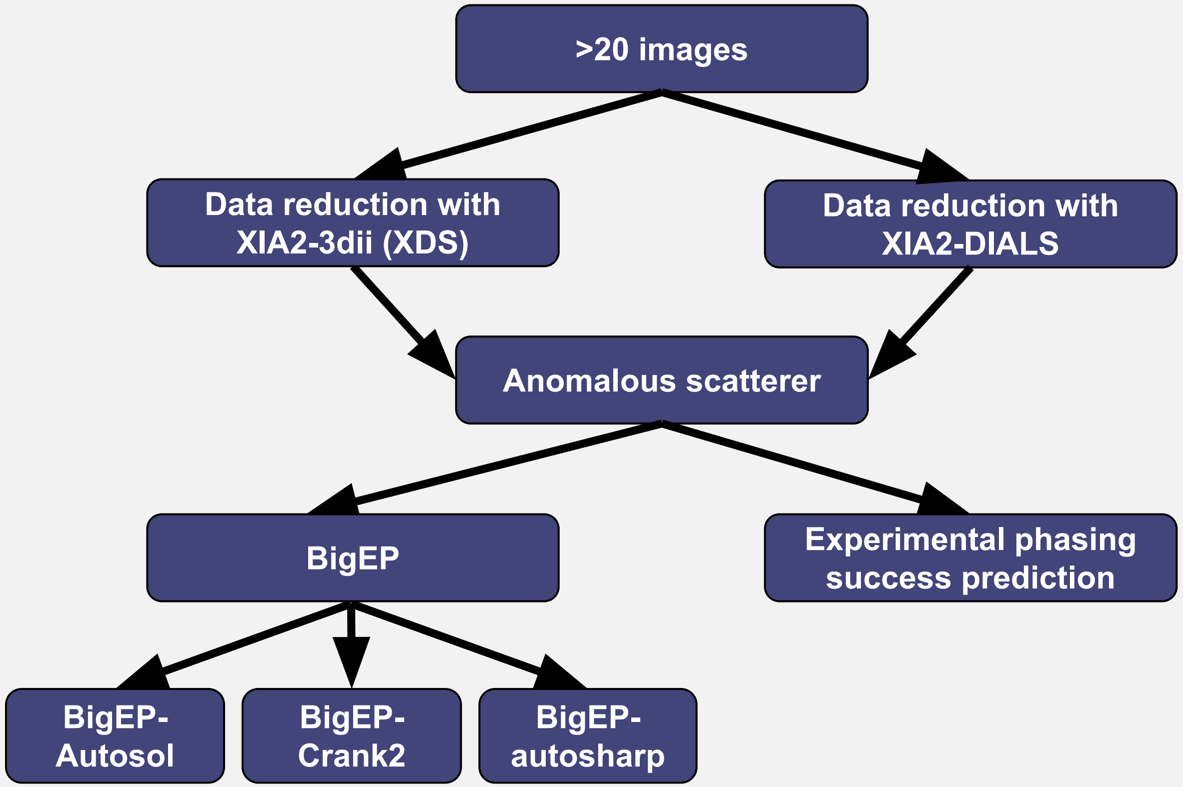
High chances of success

Low chances of success

Prediction

Raw data usage

Diamond data analysis pipelines



>20 images
Data reduction with XIA2-3dii (XDS)
Data reduction with XIA2-DIALS
Anomalous scatterer
BigEP
Experimental phasing success prediction
BigEP-Autosol
BigEP-Crank2
BigEP-autosharp

Very poor performance in run1; most samples in opposite class

Improved performance for run 2/3 after training with run1 data

3dii (XDS) – run2/3/4

- 70% of positive samples correct
- 43% of negative samples correct
- Too optimistic

DIALS – run2/3/4

- 9% of positive samples correct
- 94% of negative samples correct
- Too pessimistic

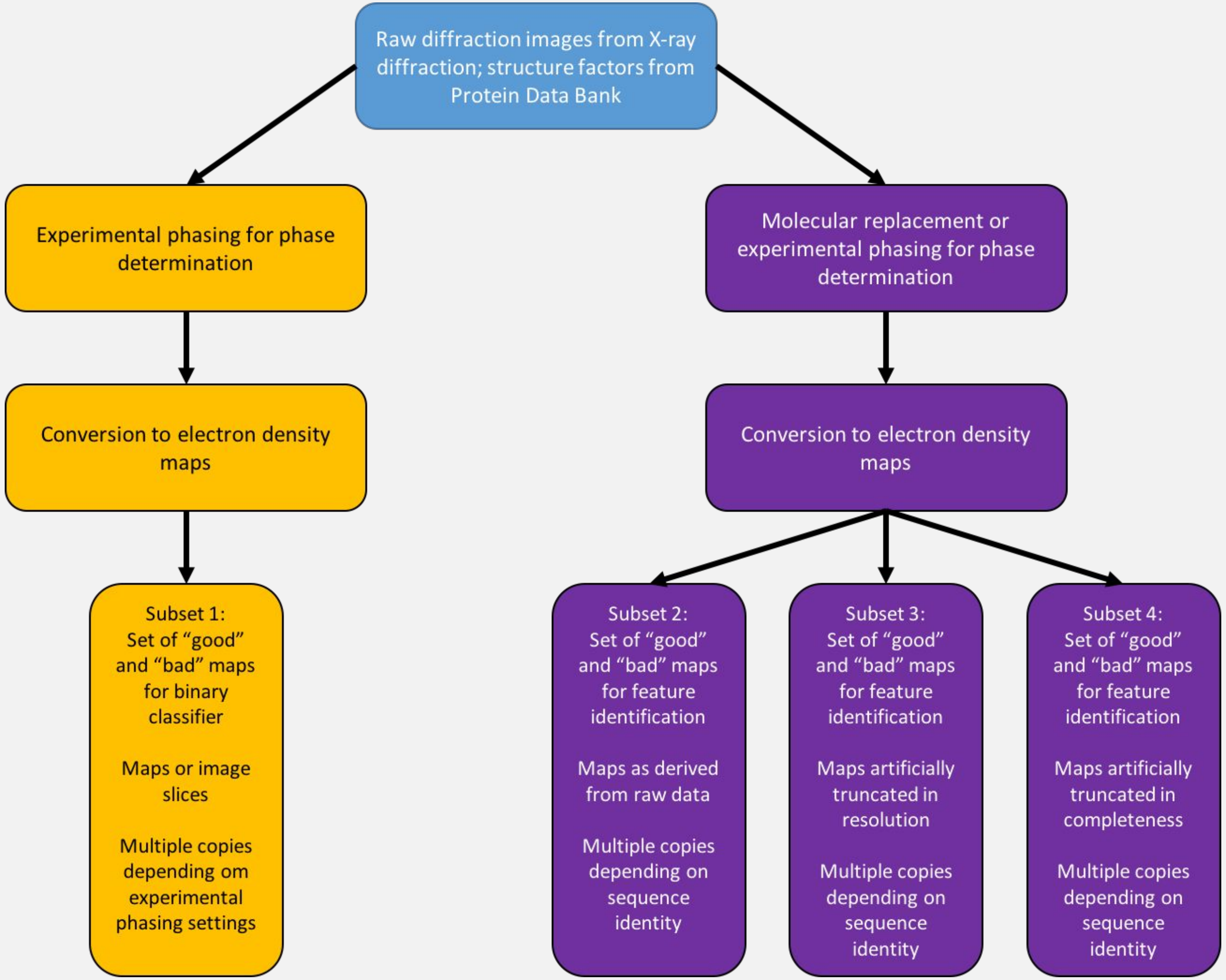
All samples have been run through BigEP pipeline; predictor itself runs on every sample with an anomalous scatterer

9-months evaluation – run1_2020 to run4_2020

	XIA2-3dii (XDS)		XIA2-DIALS	
	run1	run2/3/4	run1	run2/3/4
Class accuracy (%)	23	54	24	61
Class error (%)	77	46	76	39
Sensitivity (%)	19	70	0	9
Specificity (%)	24	43	30	94
False-positive rate (%)	76	57	70	6
Precision (%)	7	46	0	46
F1-score (%)	10	56	0	15
TP	5	284	0	19
TN	23	252	25	327
FP	71	330	58	22
FN	22	119	21	198
P	27	403	21	217
N	94	582	83	349
total	121	985	104	566

Raw data usage

Training a map quality assessor



Raw diffraction images from X-ray diffraction; structure factors from Protein Data Bank

Experimental phasing for phase determination

Conversion to electron density maps

Subset 1:
Set of “good” and “bad” maps for binary classifier

Maps or image slices

Multiple copies depending om experimental phasing settings

Molecular replacement or experimental phasing for phase determination

Conversion to electron density maps

Subset 2:
Set of “good” and “bad” maps for feature identification

Maps as derived from raw data

Multiple copies depending on sequence identity

Subset 3:
Set of “good” and “bad” maps for feature identification

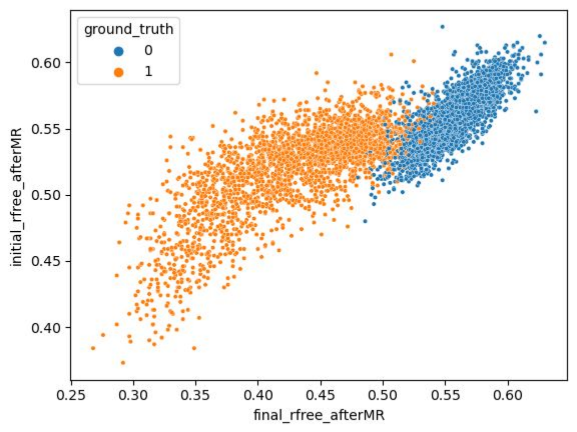
Maps artificially truncated in resolution

Multiple copies depending on sequence identity

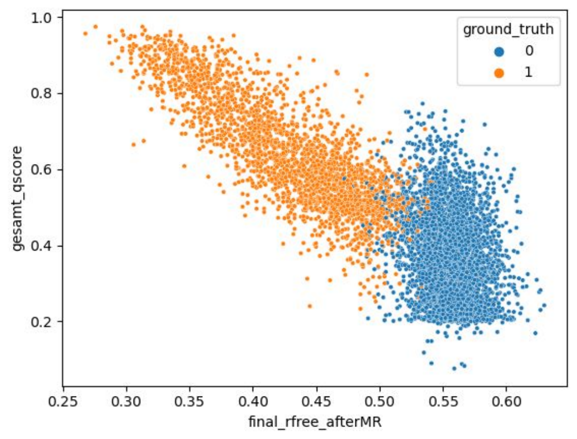
Subset 4:
Set of “good” and “bad” maps for feature identification

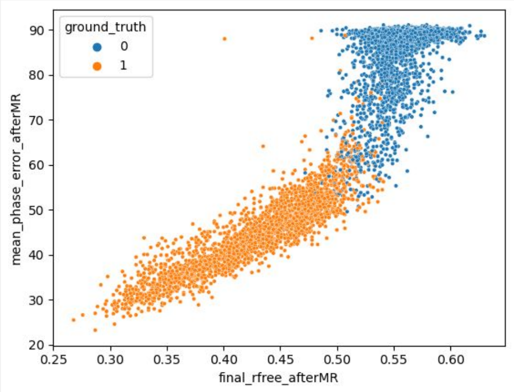
Maps artificially truncated in completeness

Multiple copies depending on sequence identity



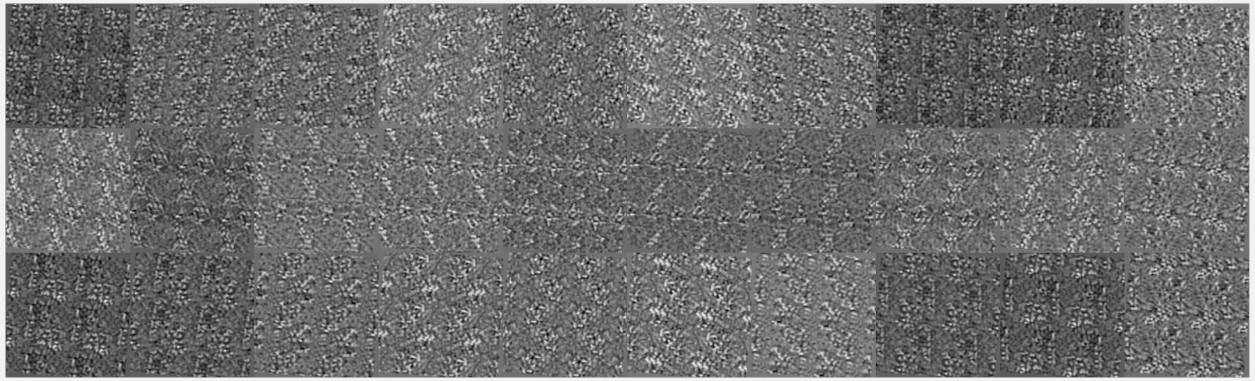
condition that 'final_rfree_refmac_default_afterMR_Buccaneer' < 0.5 for class 1



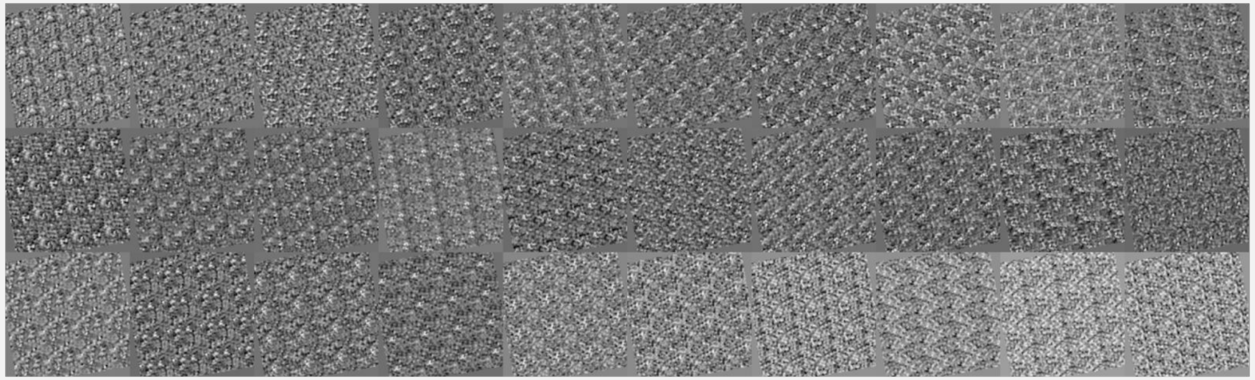


Raw data usage

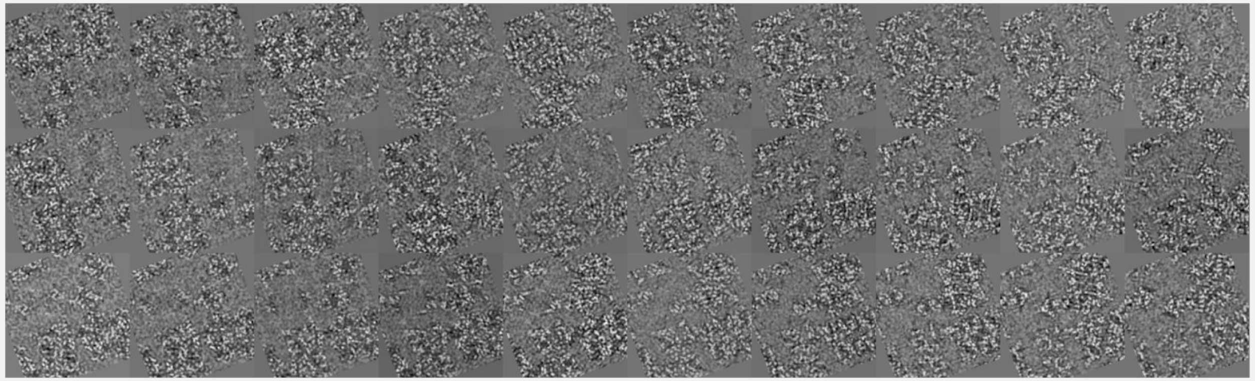
Training a map quality assessor



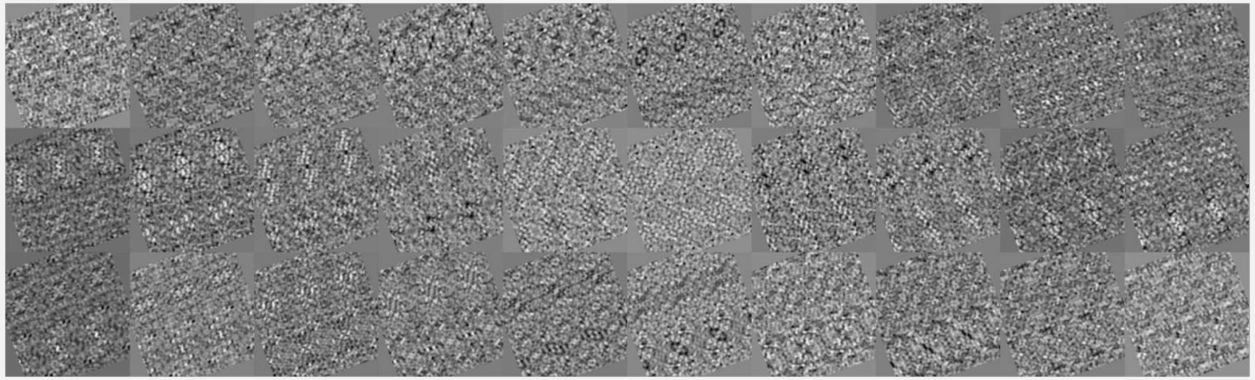
normalised map for class 1 (phased); Refmac after MR



normalised map for class 0 (not phased); Refmac after MR



normalised map for class 1 (phased); Refmac after MR



normalised map for class 0 (not phased); Refmac after MR

Acknowledgement

Diamond

James Parkhurst
Jenna Elliott (summer student 2018)
Tim Guite
Dominic Jaques (summer student 2016)
Gwyndaf Evans
Irakli Sikharulidze

MRC-LMB

Garib Murshudov
Rob Nicholls

CCP4

David Waterman
Eugene Krissinel

University of Newcastle

Arnaud Baslé

Supported by Diamond Light Source and BBSRC grants BB/L007398/1, BB/S006699/1 to GE

arise@embl.org

ARISE - Career Accelerator for Research Infrastructure Scientists

ARISE project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945405.